A Generalized Vertical Projection Histogram Using Multi-Plane Homology

Y. Yan, M. Xu and J. S. Smith

Vertical projection histograms are an efficient shape representation for 2D binary silhouettes and have been widely used in pedestrian localization for video surveillance. The weakness of this method is that it is not invariant to rotation. In this paper, a generalized vertical projection histogram is proposed to solve this problem, in which the homology transformations of the foreground silhouettes, for a set of parallel planes, are warped to, and accumulated, in the original foreground map. Then a method, similar to the vertical projection histogram, is carried out to localize pedestrians in the foreground silhouettes. This algorithm is an integrated approach using both image projection and geometric projection. Its value is demonstrated in a case study on pedestrian localization with cast shadows.

Introduction: Vertical projection histograms are a compact way to represent the shape of a 2D binary silhouette. They are computed by projecting a binary foreground region onto a horizontal axis and counting the number of foreground pixels along vertical coordinates. Such histograms have been widely used to localize multiple people in a group [1][2] and separate pedestrians from their cast shadows [3]. However, this method is not invariant to rotation. It is sensitive to skewed images or the perspective geometry in which pedestrians standing upright may look skewed in the image and converged to a vanishing point. As a solution, a generalized vertical projection histogram based on the homology transformations between multiple parallel planes is proposed.

Planar homology is a mapping between the points on a pair of parallel planes in the same camera view [4]. It is different from a homography which is the mapping of the points on the same plane in a pair of camera views. Criminisi et al. [4] used vanishing lines and vanishing points to estimate the homology. They described how errors in image measurements propagate in the homology transformation. Micusik and Pajdla [5] proposed an automatic camera calibration method using head-to-foot homology estimation from pedestrians standing at different positions in a camera view. Kilambi et al. [6] warped the foreground region of each group of people to a top view by using homographies for the ground plane and head plane, and then estimated the number of people from the intersection area. This is in spirit similar to the head-to-foot homology adequate for pedestrians of average height only.

The contributions of our work are twofold: it is a rotation-invariant vertical projection histogram; the multi-plane homology enables the localization of pedestrians of different heights. Its value is demonstrated in a case study on pedestrian localization with cast shadows.

Homology Estimation: Suppose points P_1, P_2 and P_3 are three points on the ground plane; the perpendicular lines, through points P_1, P_2 and P_3 of the ground plane, intersect a parallel plane at points P_1', P_2' and P_3', respectively, as shown in Fig. 1. The three lines P_1P_1', P_2P_2' and P_3P_3' are parallel in world coordinates and intersect at a vanishing point V in image coordinates. The three pairs of points are related by the same homology. In a homogeneous coordinate system, the image coordinate of point P is represented as $\mathbf{p} = [u, v, 1]^T$ and that of point P' at a height of h is represented as $\mathbf{p}' = [u', v', 1]^T$. The homology mapping is defined by a 3×3 transformation matrix $\mathbf{H}_h$ for corresponding points on these two planes in the image:

$$\mathbf{p}' \cong \mathbf{H}_h \mathbf{p}. \tag{1}$$

where $\cong$ denotes the equivalence defined up to scale.

The homology matrix $\mathbf{H}_h$ can be calculated after camera calibration. Let $\mathbf{p}^w = [X_w, Y_w, Z_w, 1]^T$ be a point in the world coordinates and $\mathbf{p}^c = [u, v, 1]$ be the corresponding point in the image, $\mathbf{p}^w$ and $\mathbf{p}^c$ are related by a 3×4 projection matrix $\mathbf{p}^c \cong \mathbf{M}\mathbf{p}^w = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4]\mathbf{p}^w$, where $\mathbf{m}_1$, $\mathbf{m}_2$, $\mathbf{m}_3$, and $\mathbf{m}_4$ are 3×1 vectors. The $\mathbf{M}$ matrix is constituted of the intrinsic and extrinsic parameters of the camera, which can be extracted from the camera calibration.

If the points $\mathbf{p}^w$ and $\mathbf{p}^c$ are on the ground plane, $\mathbf{p}^w$ can be seen as $\mathbf{p}_0^w = [X_w, Y_w, 0, 1]^T$. It corresponds to the point $\mathbf{p}_0^t = [u_t, v_t, 1]^T$ in a virtual top view, where $u_t = X_w$, $v_t = Y_w$. One has $\mathbf{p}_0^c \cong [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4]\mathbf{p}_0^w = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_4]\mathbf{p}_0^t = \mathbf{H}_0\mathbf{p}_0^t$ or $\mathbf{p}_0^t \cong \mathbf{H}_0^{-1}\mathbf{p}_0^c$,

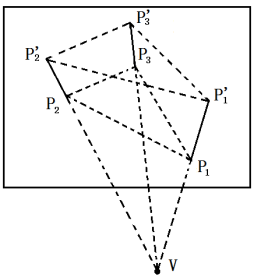

Fig. 1 *Homology mapping between a pair of parallel planes. Lines P_1P_1', P_2P_2' and P_3P_3' are perpendicular to these two planes.*

where $\mathbf{H}_0 = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_4]$ is the ground-plane homography from the virtual top view to the image.

For a plane parallel to the ground plane and at a height of h, let $\mathbf{p}_h^w = [X_w, Y_w, h, 1]^T$ be a point in the world coordinates and $\mathbf{p}_h^c$ be the corresponding point in the image. Since in the top view $\mathbf{p}_h^t = \mathbf{p}_0^t$, one has $\mathbf{p}_h^c \cong [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4]\mathbf{p}_h^w = [\mathbf{m}_1, \mathbf{m}_2, h\mathbf{m}_3 + \mathbf{m}_4]\mathbf{p}_0^t = (\mathbf{H}_0 + [\mathbf{0}|h\mathbf{m_3}])\mathbf{p}_0^t = (\mathbf{H}_0 + [\mathbf{0}|h\mathbf{m_3}])(\mathbf{H}_0^{-1}\mathbf{p}_0^c) = (\mathbf{I} + [\mathbf{0}|h\mathbf{m_3}]\mathbf{H}_0^{-1})\mathbf{p}_0^c$, where $[\mathbf{0}]$ is a 3×2 zero matrix. The homology $\mathbf{H}_h$ is then defined as:

$$\mathbf{H}_h = \mathbf{I} + [\mathbf{0}|h\mathbf{m_3}]\mathbf{H}_0^{-1}. \tag{2}$$

The foreground regions can be warped to the same image by using the homology from the ground plane to the parallel plane at height h.

Pedestrian Localization: Background subtraction based on Gaussian mixture models is used for foreground detection. The pixels of significant variation become a part of a binary foreground region map M after connected component analysis. The foreground region map M, as shown in Fig. 2(b), includes both pedestrians and background appearance changes such as cast shadows. Suppose M contains N foreground regions $R_i, i \in [1, N]$. N individual region maps are made from M and each contains one region only, e.g. M_i contains R_i only.

To prevent the overlap of different regions in accumulated homology transformations, each M_i is then warped to and accumulated in the same image, according to the homology transformations between the ground plane and each of a series of planes, $l = [1, L]$, which are parallel and at different heights $h_l, l \in [1, L]$:

$$M_i^H = M_i + \sum_{l=1}^{L} \mathbf{H}_{h_l}(M_i), \tag{3}$$

Since the feet of each pedestrian are usually within the foreground region, the multi-plane homology map M_i^H is then masked by the individual region map:

$$M_i^{HM} = M_i^H \otimes M_i. \tag{4}$$

where $\otimes$ denotes pixelwise multiplication.

Fig. 2(d) shows an example of M_i^{HM}. The most heavily overlapped areas in each foreground region are at the foot areas. However, for different foreground regions, the number of layers in the most overlapped areas are not always the same because pedestrians may have different heights. In addition, a foreground region may contain more than one pedestrian. As it is hard to find a global threshold to determine the foot areas, a staged process similar to the vertical projection histogram is applied on M_i^{HM} to find the foot points in each foreground region, as follows, where u and v are the horizontal and vertical coordinates:

1) Calculate the histogram of the maximum number of layers along the horizontal coordinates of the multi-plane homology map M_i^{HM}: $hist_i(u) = max_v M_i^{HM}(u, v)$ (see Fig. 2(e)). Potential narrow and adjacent peaks in the histogram are merged.
2) Find all the local maxima (or the mid-points of the plateaus) which reach at least half of the maximal number of layers: $u_i^j = argmax_u\{hist_i(u) \mid hist_i(u) > L/2\}$.
3) For each local maximum in the histogram, scan the vertical coordinates and find the global maximum (or the mid-point of the highest plateau) in M_i^{HM}: $v_i^j = argmax_v\{M_i^{HM}(u_i^j, v)\}$.

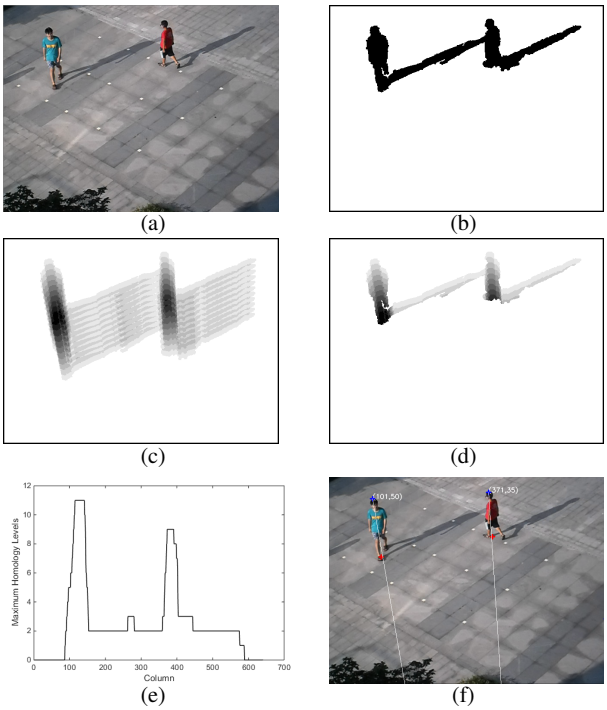

Fig. 2 *The results for pedestrian localization in cast shadows: (a) the original image, (b) the foreground region map M_i, (c) the multi-plane homology map M_i^H, (d) the masked homology map M_i^{HM}, (e) the vertical maximum projection hostogram, and (f) the detected foot points as shown in red dots. The white lines connect the manually identified tops of heads (in blue dots) and the vanishing point in the vertical direction.*

Experimental results: This algorithm has been tested over a range of video sequences. The video sequences used in this paper were captured in the authors' campus. The image size of the video sequences is 640×480 pixels and the frame rate is 15 fps. In the experiment, a ten-level homology mapping ($L = 10$) was applied to the foreground regions, which are for planes at heights of $20cm$, $40cm$, ..., and $200cm$, by using a contour-based real-time implementation [7]. Fig. 2 shows an example of the pedestrian localisation with cast shadows. In the masked homology map in Fig. 2(d), a darker colour represents more layers being overlaid. The detected foot points are shown in Fig. 2(f) as red dots. They are compared with the ground-truth positions of the feet. The ground-truth positions of the foot points were estimated as the intersection of the foreground region bottom by a line which connects the manually identified top of head and the vanishing point in the vertical direction.

Fig. 3 illustrates a comparison between the existing vertical projection histogram and the proposed algorithm. In a real-world scenario, due to a camera skew and the perspective geometry, pedestrians standing upright on the ground may not always appear upright in an image. In order to evaluate the performance of these two algorithms, the original frames were rotated $15°$ anticlockwise (Fig. 3(a)) to simulate an image captured with camera skew. Fig. 3(b) and Fig. 3(c) are the foreground region map and the masked homology map respectively. Fig. 3(d) shows the vertical projection histogram and Fig. 3(e) shows the vertical maximum projection histogram. The lines in Fig. 3(b) and Fig. 3(c) indicate the global maxima or the central position of the highest plateau in Fig. 3(d) and Fig. 3(e), respectively. Table 1 shows a statistical result of the comparison. The original frames were not rotated, rotated $15°$ anticlockwise or rotated $25°$ anticlockwise, respectively. The results show the mean and standard deviation of the localization errors between the detected foot points and the manually measured ground truth. Compared with the vertical projection histogram, the proposed method is more robust to image rotation.

Conclusions: A generalized vertical projection histogram is proposed to implement a rotation-invariant shape representation. In this method, the homology transformations of the foregrounds are projected to and accumulated in the foreground map, and then the vertical maximum projection histogram is generated to localize pedestrians in the original foreground regions. This method can be used for either the localization of multiple people in a group or the separation of people from cast shadows. The experimental results have shown its robustness in rotation invariance.

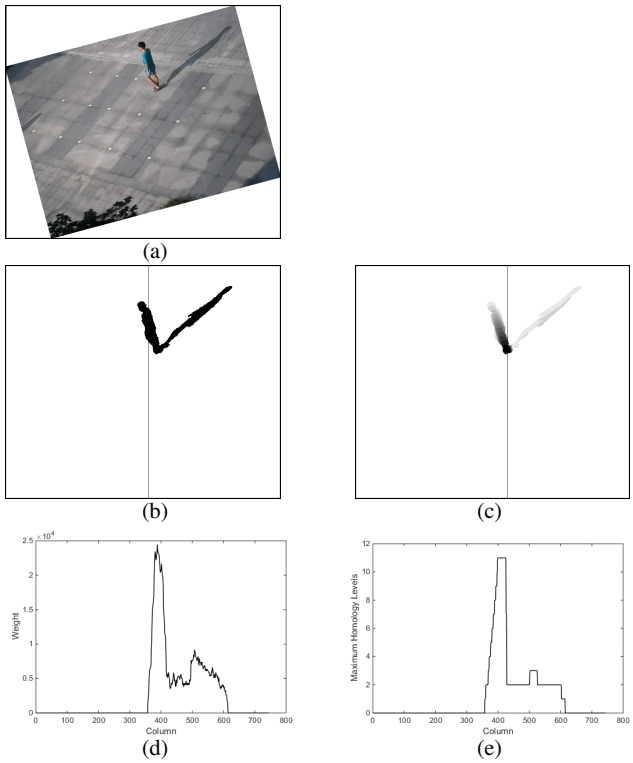

Fig. 3 *A comparison of the vertical projection histogram and the proposed method: (a) the rotated frame by 15 degrees anticlockwise, (b) the foreground region map, (c) the masked homology map, (d) the vertical projection histogram of (b), and (e) the vertical maximum projection histogram of (c). The detected foot points are along the vertical lines in (b) and (c).*

Table 1: Localization errors of foot points on rotated frames.

Degree of rotation	Errors of vertical projection histogram (pixels)	Errors of the proposed algorithm (pixels)
$0°$	15.5 ± 6.4	3.1 ± 2.1
$15°$	24.1 ± 9.6	3.2 ± 2.0
$25°$	41.1 ± 17.5	4.5 ± 2.7

Acknowledgment: This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 60975082 and an XJTLU PhD scholarship under Grant PGRS-12-02-07.

Y. Yan and M. Xu (*Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China*) Y. Yan and J. S. Smith (*Department of Electrical Engineering and Electronics, University of Liverpool, L69 3BX, Liverpool*)

E-mail: ming.xu@xjtlu.edu.cn

References

1 Haritaoglu I. A., Harwood D. and Davis L. S.: 'W^4: Real-time surveillance of people and their activities', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (8), pp. 809-830
2 Hu W., Hu M., Zhou X., Tan T., Lou J. and Maybank S.: 'Principal axis-based correspondence between multiple cameras for people tracking', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (4), pp. 663-671
3 Hsieh J. W., Hu W. F., Chang C. J. and Chen Y. S.: 'Shadow elimination for effective moving object detection by Gaussian shadow modeling', *Image and Vision Computing*, 2003, **21**, (6), pp. 505-516
4 Criminisi A., Reid I., and Zisserman A.: 'Single view metrology', *Int. J. of Computer Vision*, 1999, **40**, (2), pp. 123-148
5 Micusik B. and Pajdla T.: 'Simultaneous surveillance camera calibration and foot-head homology estimation from human detections', *IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 1562-1569
6 Kilambi P., Ribnick E., Joshi A. J., Masoud O. and Papanikolopoulos N.: 'Estimating pedestrian counts in groups', *Computer Vision and Image Understanding*, 2008, **110**, (1), pp. 43-59
7 Xu, M., Ren J., Chen D., Smith J. S. and Wang, G.: 'Real-time detection via homography mapping of foreground polygons from multiple cameras', *IEEE Int. Conf. on Image Processing*, 2011, pp. 3593-3596